DOCUMENT RESUME

ED 453 272 TM 032 809

AUTHOR Floreck, Lisa M.; De Champlain, Andre F.; Kaplan, David
TITLE Assessing Sources of Score Variability in a Multi-Site Medical Performance Assessment: An Application of Hierarchical Linear Modeling.
PUB DATE 2001-04-13
NOTE 24p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Clinical Experience; Computer Software; Licensing Examinations (Professions); Patients; *Performance Based Assessment; *Physicians; *Reliability; *Scores; Skill Development
IDENTIFIERS *Hierarchical Linear Modeling; Multilevel Analysis; *Standardized Patients

ABSTRACT

The purpose of the current study was to use multilevel modeling to quantify and explain the sources of score variation in standardized patient (SP) encounters. Through laypersons trained to portray SPs and record medical student actions, SP examinations allow the measurement of examinees' clinical and interpersonal skills. In this study, the SP test assesses the clinical skills of physicians about to enter supervised practice. Four cases were drawn from the SP bank. The number of examinees who saw each of these cases ranged from 357 to 565 with this number being reduced in the checklist (objectively scored) models to those who had already taken step 2 of the U.S. Medical Licensing Examination. The multilevel modeling software package HLM5 was used to estimate the proportion of score variation between SPs and training sites, assess the relationship between the skill scores and SP characteristics, and quantify the proportion of variation explained when SP characteristics are added into the model. Results of previous generalizability analyses have demonstrated that there is little variation in scores across SPs and sites, and that most of the variation in scores is due to case specificity. The findings from this study show that although SP variability was negligible for checklist scores, variability was quite large for interpersonal scores. Although a major advantage of using multilevel modeling is to explain variation at various levels, the variables used in this study were not helpful in explaining the variation between SPs. Overall, however, this study should be seen as an important first step in using multilevel modeling to explore the variability of SP examinations. Careful consideration of SP and site characteristics should be captured and analyzed statistically so that steps can be taken to implement fair and reliable examinations. (Contains 4 tables and 17 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 272

Assessing Sources of Score Variability in a Multi-Site Medical Performance Assessment: An Application of Hierarchical Linear Modeling

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
L. Floreck
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Lisa M. Floreck and Andre F. De Champlain

National Board of Medical Examiners

David Kaplan

University of Delaware

TM032809

Running Head: Modeling variability in SP exams

Paper Presented at the National Council on Measurement in Education, Seattle, WA

April, 13, 2001

BEST COPY AVAILABLE

Standardized Patient (SP) Examinations are widely used by medical schools, testing and certification organizations to evaluate sets of skills not readily measurable with written multiple-choice examinations (Reznick, 2000; Whalen, 2000). Through using laypersons trained to portray SPs and record student actions, these examinations allow the measurement of examinees' clinical and interpersonal skills. Albeit valuable, these examinations bear limitations, mainly decreased reliability of examinee scores attributable to variation in SP portrayal, scoring and the limited number of cases seen by the student.

Regardless of whether SP exams are being used by a medical school for teaching purposes or a medical testing organization for licensure or certification, it is critical that scores accurately reflect the appropriate clinical skill level of the examinees. Threats to reliability may increase when exams are administered on a large scale and it becomes necessary to train multiple SPs to portray the same case across multiple testing sites. Much research has focused on quantifying sources of variability in SP exams since any type of unwanted variation could have a deleterious impact on pass/fail decisions. These studies' conclusions are not easily discerned.

The majority of initial studies indicated that the use of multiple SPs did not cause large discrepancies in total test score when examinees were randomly assigned to SPs (van der Vleuten & Swanson, 1990). Swanson & Norcini (1989) found that raters nested within a case explained only 1% to 2% of the observed score variance and De Champlain *et al.* (1998) found that multiple SPs could similarly assess examinee performance, leading to identical mastery-level decisions for nearly all students tested. Previous research has indicated that it is not variation in

raters but rather case content, i.e. case specificity, which contributes most to variability in a student's scores. (Klass, Fletcher, King, Durinzi, Nungester, Clauser & Ripkey, 1992; Swanson & Norcini, 1989; Tamblyn, Klass, Schnabl & Kopelow, 1991; van der Vleuten & Swanson, 1990).

However, other studies have indicated that multiple SPs may introduce enough error to be consequential at the case level. Swanson and Norcini (1989) found that, although the use of multiple SPs playing the same case for different examinees did not affect the total test score, there were cases in which raters disagreed. In addition, Colliver, Robbs & Vu (1991) reported statistically significant differences between failure rates among SPs simulating the same cases. More recently, differences in intra-rater reliability (DeChamplain, Macmillan, Klass, Margolis, 1999) and the effects of rater discrepencies on pass/fail decisions for heterogeneous groups have been large enough to warrant concern (DeChamplain, Gessaroli & Floreck, 2000).

Research has also shown that the use of multiple SPs and raters appears to have less influence on objectively scored measures such as checklists and more effect on the variability of interpersonal skills scores. For example, Colliver *et al.* (1994) found that the use of multiple raters on the same case decreased inter-case reliability more on measures of interpersonal and communication skills than checklists, total scores and written scores. Boulet, *et al.* (1998) found inconsistencies across holistic scoring of post encounter notes which supports previous research (Colliver *et al.*, 1994) suggesting that subjective scores are more highly influenced by individual variability. In summary, the use of multiple SPs to portray and score a given case does not impact all clinical scenarios in a consistent fashion.

Although testing organizations that accommodate large volumes of examinees are concerned with the effects of administering forms across multiple sites, fewer studies have examined these effects. While some have found little or no differences in scores of candidates taking the same test administered at different sites (DeChamplain, Macmillan, Klass, Margolis, 1999; Reznick *et al.*. 1993) others have indicated that candidates scores can be influenced by the site at which they take an exam, especially when training offered to SPs is minimal (Tamblyn *et al.*, 1991; Petrusa *et al.*, 1991). Interestingly, Tamblyn *et al.* (1991) reported a great deal of variation in the reliability of individual raters suggesting that rater characteristics may be responsible for this variability. Unfortunately, fewer studies have systematically examined the impact of rater characteristics on score variability.

It is clear that more research is needed to assess the sources of variation present when multiple SPs portray and score identical cases across multiple testing sites. According to De Champlain *et al.* (1998), we should not conclude that relatively small amounts of rater and site variation in generalizability studies are necessarily synonymous with negligible effects on examinee scores. Rater components are disproportionately small because the variation associated with case content is typically very large. The effect on mastery level decisions and rank ordering of examinees could very well be affected.

One limitation of (the commonly used) generalizability theory in quantifying sources of score variation in SP exams is that crossed designs are more desirable than nested designs (Shavelson & Webb, 1991). When nesting is inherent in the design many variance components

cannot be estimated. A further limitation is that generalizability analyses do not allow us to address other important issues including characteristics that contribute to unwanted sources of variation. Understanding sources of score variation is helpful but this alone cannot help test developers to reduce unwanted variation unless actual *causes of variation* are known. Multi-level modeling makes it possible to not only quantify sources of variation which would be difficult to estimate using generalizability analysis, but to examine how factors such as SP gender and experience explain such variation.

The purpose of the current investigation was to use multi-level modeling to quantify and explain, in a more comprehensive manner, the sources of score variation in SP encounters. The partitioning of variance and covariance components among various levels (e.g. student, rater, test site) and determining the relative weight and significance of individual SP characteristics will provide important information to eventually implement a fair and reliable SP exam. Additional information on sources of score variability will enable us to make more informed decisions regarding scoring, calibrating and equating procedures that will ultimately enhance decision consistency and accuracy rates. The models selected for this investigation allow for estimation of rater and site effects without the large variance components related to case specificity which provides important feedback for case development and training activities.

Method

SP examination and measurement instruments

In the present study, the SP test assesses the clinical skills (history taking, physical examination, communication) of physicians about to enter supervised practice. Examinees

proceed through (six to ten) cases and encounter patients in a setting intended to reflect an ambulatory clinic. Subsequent to each 15-minute encounter the SP performing the case records the performance of students using a checklist and the Patient Perception Questionnaire (PPQ).The checklist is tailored specifically to the standardized patient's complaint by a group of subject matter experts and contains 10 to 25 dichotomously scored items targeting behaviors deemed critical for success on the encounter. Unlike the checklist, the case-invariant PPQ is comprised of seven 5-point Likert type items that measure the student's interpersonal skills (IPS). Percent correct scores are calculated for both checklist and IPS scores. The reliability (Cronbach's alpha) of checklist and IPS scores is typically lower than traditional multiple-choice examinations due to the limited number of items.

One final measure, USMLE™ Step 2, was used to adjust for ability when modeling checklist scores. According to Bryk & Raudenbush (1992), the use of a covariate related to the dependent measure (in multi-level modeling) is useful because it reduces the unexplained variance at level-1 and increases precision of estimates at higher levels. Descriptions of variables are provided in Table 1 and summary descriptive measures are shown in Table 2.

Methodology

Four cases were drawn from the bank. One case measured biomedical skills, another grave illness and two measured routine counseling skills. The second routine counseling case was performed by both a male and female SP whereas all other cases were performed by SPs of identical gender. These cases had been administered with variable frequency across testing sites in 2000. The number of examinees who saw each of these cases ranged from 357 to 565 with

this number being further reduced in the checklist models to those who had already taken USMLE Step 2.

The multi-level modeling software package, HLM5 (Bryk, Raudenbush & Congdon, 1994) was used to (1) estimate the proportion of score variation between SPs and training sites (2) assess the relationship between the skill scores (checklist or IPS scores) and SP characteristics and (3) quantify the proportion of variation explained when SP characteristics are added into the model. Each of the seven skill scores (3 checklist and 4 IPS) was modeled separately since prior research has reported that IPS scores are more prone to SP variation. Skill scores were also modeled as a function of the number of encounters performed by the SP over the course of the testing period and gender (solely for the second routine counseling case portrayed by both a male and female SP). The number of encounters served as a proxy for test administration experience. USMLE Step 2 was used as a covariate for the checklist models since it was moderately correlated with SP checklist scores. However, since the relationship with IPS scores was weak, these were run as intercept-only models with no covariate.

Modeling Skill Scores in SP Examinations

A 3-level one-way ANOVA with random effects was run to estimate variation (1) among students encountering a given SP (2) among SPs at a given site and (3) across sites. Predictors were entered in the ANOVA in a block entry fashion. First, USMLE Step 2 scores were entered at the student level (Level 1) to adjust for ability (for checklist scores only). The number of SP encounters was entered at the SP level (Level 2) to help explain variation between SPs at a given site. Gender was also added as a SP level predictor for the routine counseling case performed by

both a male and a female SP. Due to the limited sample sizes, none of the models included test site predictors (at Level 3). Random effects which were not statistically significant were removed from the models to reduce the number of parameters. Models with no variation among testing sites were reduced to 2-level models prior to adding predictors. Finally, models with little variation among SPs or testing sites were not modeled with predictors. More detail is provided below.

Model 1-One way ANOVA

Prior to entering predictors into the model, the one-way ANOVA with random effects was run to estimate the differences in means at each level. This is an important step to undertake because it provides a point estimate of the grand mean and is necessary to measure the variation explained when predictors are entered in subsequent models. Using this model, intra-class correlations were calculated to determine the proportion of variance in skill scores attributable to training site and standardized patients. The ANOVA models and those including predictors are provided in Table 3. The parameters are interpreted as follows:

γ_{000} Grand mean (Checklist or IPS);

β_{00k} The mean in site *k* for SPs;

π_{0jk} The mean for SP *j* in site *k*;

u_{00k} The deviation in site *k*'s mean from the grand mean (training site effect);

r_{0jk} The deviation in SP *j*'s mean from site *k*'s mean (SP effect);

e_{ijk} The deviation in student *ijk*'s score from his/her SP's mean (student effect).

Checklist Scores

Equation 1 describes the level-1 model for checklist scores whereby centering the covariate around the grand mean produces an intercept adjusted for student ability.

$$\text{CHECKLIST} = \pi_{0jk} + \pi_{1jk}(\text{Step } 2_{ij} - \text{Step } 2..) + e_{ijk} \quad (1)$$

Thus, π_{0jk} becomes the mean checklist score for SP_j in site k after adjusting for Step 2. Similarly, the variance at the student level, e_{ijk} is the residual variance after adjusting for Step 2. We assume this residual variation to be independent and normally distributed. The slope coefficient, π_{1jk}, reflects the number of Step 2 score points required to increase the checklist score by 1 percent. At this stage the level-2 and level-3 models are both unconditional (contain no predictors).

Subsequent to adding Step 2, the number of SP encounters (#SPenc) was entered to predict variation in SPs. This 2-level model is shown below in equation 2. Since #SPenc is

$$\pi_{0jk} = \beta_{00k} + \beta_{01}(\#SPenc_j - \#SPenc..) + r_{0jk} \quad (2)$$

centered around the grand mean, β_{01} reflects the number of SP encounters required to produce a 1 point increase in the checklist score for a student of average ability. Whereas π_{0jk} is the Step 2-adjusted checklist score for SPj in site k, β_{00k} is the Step 2-adjusted checklist score in *site k* further adjusted for #SPenc. Level-3 is left unconditional, i.e. we are not attempting to use predictors at the training site level to model SP variation. This series of checklist score models is delineated in Table 3a.

IPS scores

The IPS scores were modeled similarly to checklist scores with the exception of a level-1 covariate which was excluded. The random intercept model including the SP predictor (#SPenc) is provided in Table 3b (model 2). Note that the interpretation of β_{01} differs without the covariate and now reflects the number of SP encounters that are required to produce a 1 point increase in IPS scores without having conditioned on ability.

The routine counseling case performed by a male and female SP was modeled differently from the other models (see Table 3c). A 3-level model was not utilized because there were not data from multiple SPs for many of the sites. Another difference is that checklist scores were not modeled due to lack of Step 2 data. Unlike previous models, the gender indicator variable (Female) was entered prior to #SPenc. Equation 3 shows the 2-level model in which the intercept, γ_{00}, represents the average IPS score for a male who has seen the case by a SP with average experience.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Female}) + \gamma_{02}(\#SPenc - \#SPenc..) + u_{0j} \qquad (3)$$

The gender coefficient, γ_{01} reflects the difference in means between female and male SPs and γ_{02} represents the number of encounters required for the IPS score to increase by 1 point. The intercept, γ_{00}, is the IPS score expected to be given by a male SP with average experience.

Results

Proportion of Score Variation among SPs and Training Sites (ANOVA)

Results from the seven ANOVA models indicated that checklists were more robust to variation among SPs within a given site (2%-3%) whereas IPS scores varied substantially across SPs in a given site (26%-49%). Irrespective of the nature of the case, less than 12% of the variation in checklist scores was due to differences across testing sites. Checklist scores for the biomedical case displayed the most inter-site variability (11.6%), whereas the grave illness case varied the least (3%) among sites. There was no variation in interpersonal skill scores across testing sites with the exception of one routine counseling cases where 16% of the variation in IPS scores was attributable to training site. The findings suggested that multilevel modeling was appropriate for all skill scores except the communication checklist score. ANOVA results are presented in Table 4.

Checklist scores – Routine counseling encounter

The variance components estimated in the one-way ANOVA (Table 4) show that 8.3% of the variation in scores is due to site differences while only 2.7% is due to SP differences. After conditioning on ability, the mean checklist scores still varied across training sites (u_{00k}, χ^2=29.06, df=8), indicating that it is possible to explain and reduce this variation. When #SPenc was added to the model the conditional variance component, u_{00k}, (representing the variability in the grand mean after controlling for the number of SP encounters and Step 2) was still statistically significant (u_{00k}, χ^2=32.52, *df*=8). As mentioned previously, no site predictor variables were available to model site variation in the current investigation.

Regarding SP variation, we reject the null hypothesis and conclude that there are differences between SPs after controlling for the number of SP encounters and ability (r_{0jk}, χ^2=18.79, *df*=9). Although Step 2 was statistically significant, the number of SP encounters was not statistically related to checklist scores. The variance explained by the inclusion of this predictor (23%) is misleading because there is little variation among SPs to begin with. Consequently, a reduction by 23% amounts to less than 1% of the variation in total scores. The deviance statistics indicate that the addition of #SPenc is not justified over the more parsimonious model containing only Step 2 (χ^2 = 0.94, *df*=1). Due to the unique modifications made to each of the models, random coefficients are not provided in table 4.

Checklist scores – Biomedical encounter

The results of the 3-level ANOVA in Table 4 show that approximately 11.6% of the variation in checklist scores was due to training site differences and that only 2.9% was due to differences among SPs. This model indicates that after adjusting for ability, students who encounter a SP with average experience (#SPenc) will have a checklist score of approximately 53%. Their score is estimated to increase by 1 for every 6 points they achieve on the Step 2 examination (over and above the mean Step 2 score for this sample). Similar to findings reported for the routine counseling case, #SPenc was not significantly related to checklist score, most likely due to the small proportion of variation left to explain. The addition of #SPenc is therefore not justifiable (difference in deviance between Step 2 model and #SPenc model = 0.10 with 1 degree of freedom). The estimated reliability of the intercept is moderately high in these models (.72 to .77) suggesting these data provide an adequate level of precision in estimating β_{00k} from

the current sample.

IPS scores - Biomedical Encounter

SP predictors were added to a 2-level model because there was no variation across training sites. This proxy for test administration experience was not significantly related (γ_{01}= -.05, t= -0.40) to IPS score variation. Consequently, modeling IPS scores using this SP characteristic did not explain variation among SPs. The significance of the SP effect (u_{0j}, χ^2=380.19 df=25) indicates that after adding this predictor, the mean IPS scores still varied around the grand mean. Had the χ^2 statistic not been statistically significant we would have concluded that there was no variation among SP means. However, we failed to reject the null hypothesis of homogeneity.

The difference in deviance statistics between model 1 and model 2 was distributed approximately as χ^2=0.00 with 1 degree of freedom which indicates that the addition of #SPenc was not justified. In summary, the ANOVA model was most useful in describing the variation in these scores: 26.5% at the SP level and 0% at the training site level. Other SP characteristics are necessary to decompose sources of variation in IPS scores for biomedical encounters.

IPS scores – Grave illness encounter

The results for this series of analyses mirror those reported above. Again, the number of SP encounters was not useful (γ_{01} = -0.04, t= -0.24) in explaining differences among SPs. This model did not explain any of the variance at the SP level (44.8%). Consequently, there remains a great deal of unexplained variation, (u_{0j}, χ^2=376.30, df=21). The addition of #SPenc is not justified over the 3-level ANOVA.

IPS Scores – Routine Counseling 1

The results provided in Table 4 indicate that 27.1% of the variation in IPS scores was related to SP differences. Similar to the biomedical encounter, a 3-level model was not appropriate since IPS scores did not vary across testing sites. Adding the number of encounters (#SPenc) into the 2-level model explained 56.4% of this variation among SPs although this predictor was not statistically related to IPS scores. The results of these scores differ from previous findings but still do not indicate that test administration experience plays a role in score variability.

IPS Scores – Routine Counseling 2 (male & female SP)

IPS scores for the routine counseling case portrayed by both a male and female were also not influenced by test administration experience. In fact, the results from table 4 show that neither gender ($\gamma_{01} = 4.08$, $t = 0.72$) nor #SPenc ($\gamma_{02} = 0.13$, $t = 0.82$) was statistically related to IPS. Adding these predictors decreased the intercept from 78.12 (ANOVA) to 75.88 (Female) and finally to 72.69 (Female & #SPenc). However, the variation in γ_{00} decreased by only 3.1% in Model 2 and 7.6% in the final model. Neither model was justified.

Discussion

This investigation has provided important information for the consortia that develop SP tests to be administered across multiple testing sites and within a given site. Results of previous generalizability analyses have demonstrated that there is little variation in scores across SPs and sites, and that most of the variation in scores is due to case specificity, i.e. the nature of the complaint. The findings from this study show us that although SP variability was negligible for

checklist scores, variability was quite large for interpersonal skills scores. This result is supported by previous research indicating that interpersonal skills score are more influenced by variations among SPs. Based on this finding it is important to consider how these instruments are developed and to do so in such a way that limits rater subjectivity. For example, asking the rater to record how he or she 'felt' may not be adequate. Rather, videotaped encounters that establish baselines of interpersonal skill levels may be more effective.

One result not anticipated based on past research was the systematic SP stringency/leniency across training sites for checklist scores of certain cases. After adjusting for ability, several checklist scores varied by as much as 10% as a function of the testing site. This inter-site variability in checklist scores could have resulted from differences in training across sites or possibly because guidelines to scoring checklists were not clear (i.e. what a student must do to receive credit for completing a given behavior). This might also be due to the various levels of adherence to protocols on the part of different trainers. Although past research has shown the effect of testing site to be minimal at the overall test level, this study underscores the importance of looking at cases on an individual basis. If systematic stringency or leniency among SPs or sites is detected during pre-tesing, steps could be taken prior to live administration to remedy the situation. Unlike checklist score, IPS scores did not differ as a function of testing site, with the exception of one routine counseling case.

Although a major advantage of using multilevel modeling is to *explain* variation at various levels, the variables utilized in the current investigation were not helpful in explaining the variation between SPs. The benefit of adding predictors into multilevel models diminishes if

there is not a substantial amount of variation left to model (generally about 10%). Therefore, based on the negligible amount of SP variability in the checklist scores it is doubtful that test administration experience or any predictor, for that matter, could have explained this variation. It was surprising, however, that the number of encounters performed by the SP had little impact in explaining variation in IPS scores. The aggregated nature of this variable may have contributed to the lack of significance. It is also possible that results might have differed had an adequate covariate been used to adjust for ability on the IPS scores.

This preliminary study has several limitations that must be addressed. As mentioned above, the results from IPS models must be interpreted cautiously because the model lacked an adequate covariate. However, in a 3-level model the proportion of variation among SPs is associated with those performing at a given site. Under the assumption that students at a given school are of the same relative ability, we would not anticipate the proportion of variation estimated in the ANOVA to be so large. A second limitation is related to the limited nature of the data set. A more expansive data set containing extensive student, SP and site related information would be desirable in order to explore additional sources of SP and site variation. A final limitation is that for the most part, only one case was selected to represent the nature of encounters. It is important to note that using HLM at the case level precludes the estimation of SP or site variation across the entire examination (as is estimated using generalizability analysis). The methodology described in this study is more desirable for pre-testing cases than estimating variance proportions at the exam level. Since methodology should be selected to answer the research question at hand, this is not seen as a limitation.

Unlike generalizability analysis, HLM models are able to distinguish which factors of SPs or sites may be responsible for score variation. This information is critical to large scale testing organizations that are considering selection of equating, calibrating and scoring procedures. If the sources of score variation are known, then steps can be taken *a priori* to reduce this variation and scoring and calibrating routines may be developed to compensate for variability in the measures.

Overall, this study should be viewed as an important first step in utilizing multi-level modeling to explore the variability of SP examinations. Some results from this investigation confirmed those found in previous research whereas other findings offered a different perspective. Given the high stakes involved in a licensing examination, it appears unwise to adopt the attitude that stringency and leniency of SPs "wash out" over the course of the (multi-case) examination. Given the opportunity, it is advisable to take actions prior to live administration to reduce the variation in individual cases. In addition to quantifying the score variation, HLM models inform us as to the causes of variation inherent to clinical skills encounters. Careful consideration of SP and site characteristics should be captured and analyzed statistically so that steps can be taken to develop and implement fair and reliable examinations.

TABLE 1: Variable names, Descriptions, Scales

NAME	DESCRIPTION	SCALE
LEVEL-1		
CHECKLST	Instrument measures a composite of History Taking, Communication and Physical Exam Skills	Percent Correct (1-100%)
IPS	Patient Perception Questionnaire of Interpersonal Skills	Percent Correct (1-100%)
STEP 2	USMLE Step 2 score	Scale has mean of 220 & s.d of 22
LEVEL-2		
#SPENC	Number of encounters performed by the SP	Interval
FEMALE	Dummy coded predictor variable for gender	Females are coded 1
LEVEL-3		
NONE		

TABLE 2: Descriptive Statistics

Encounter	N	Mean	Standard Deviation	Minimum	Maximum
Grave Illness					
Checklist	519	76.22	14.55	22	100
Interpersonal Skills	526	75.60	16.09	20	100
#SPenc	23	22.87	14.33	5	64
Routine Counseling (#1)					
Checklist	538	70.42	12.40	29	100
Interpersonal Skills	543	76.50	13.25	34	100
Step 2	171	215.83	19.86	166	269
#Spenc	19	16.89	7.45	5	30
Routine Counseling (#2)					
Checklist	327	62.24	15.64	8	100
Interpersonal Skills	330	79.66	15.79	31	100
#Spenc	15	22.00	17.21	8	67
Female	15	0.60	0.51	0	1
Biomedical					
Checklist	565	50.93	14.54	14	86
Interpersonal Skills	580	79.30	12.87	37	100
Step 2	190	214.52	23.59	141	270
#SPenc	19	22.74	16.14	8	75

Table 3a: 3-Level Hierarchical Models for Checklist Data

Model	Level	Equations
1 (ANOVA)	1	$CHECKLIST=\pi_{0jk} + e_{ijk}$
	2	$\pi_{0jk} = \beta_{00k} + r_{0jk}$
	3	$\beta_{00k} = \gamma_{000} + u_{00k}$
2	1	$CHECKLIST = \pi_{0jk} + \pi_{1jk}(Step\ 2) + e_{ijk}$
	2	$\pi_{0j\,k} = \beta_{00\,k} + r_{0jk}$
		$\pi_{1j\,k} = \beta_{10k} + r_{1jk}$
	3	$\beta_{00k} = \gamma_{000} + u_{00k}$
		$\beta_{10k} = \gamma_{100} + u_{10k}$
3	1	$CHECKLIST = \pi_{0jk} + \pi_{1jk}(Step\ 2) + e_{ijk}$
	2	$\pi_{0j\,k} = \beta_{00\,k} + \beta_{01k}(\#SPenc) + r_{0jk}$
		$\pi_{1j\,k} = \beta_{10k} + r_{1jk}$
	3	$\beta_{00k} = \gamma_{000} + u_{00k}$
		$\beta_{01k} = \gamma_{010} + u_{01k}$
		$\beta_{10k} = \gamma_{100} + u_{10k}$

*italicized variable are centered around the grand mean

Table 3b: 3-Level Random-Intercept Models for IPS Data

Model	Level	Equations
1 (ANOVA)	1	$IPS = \pi_{0jk} + e_{ijk}$
	2	$\pi_{0jk} = \beta_{00k} + r_{0jk}$
	3	$\beta_{00k} = \gamma_{000} + u_{00k}$
2	1	$IPS = \pi_{0jk} + e_{ijk}$
	2	$\pi_{0j\,k} = \beta_{00\,k} + \beta_{01k}(\#SPenc) + r_{0jk}$
	3	$\beta_{00k} = \gamma_{000} + u_{00k}$
		$\beta_{01k} = \gamma_{010} + u_{01k}$

*italicized variable are centered around the grand mean

Table 3c: 2-Level Random-Intercept Models for IPS data

Model	Level	Equations
1 (ANOVA)	1	$IPS = \beta_{0j} + r_{ij}$
	2	$\beta_{0j} = \gamma_{00} + u_{0j}$
2	1	$IPS = \beta_{0j} + r_{ij}$
	2	$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Female}) + u_{0j}$
3	1	$IPS = \beta_{0j} + r_{ij}$
	2	$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Female}) + \gamma_{02}(\#SPenc) + u_{0j}$

Table 4. Variance Proportions Estimated in the one-way-ANOVA with Random Effects Model. Effect Sizes and Variation Explained for the Final Model (Including SP Predictors)

	Variance Proportions in ANOVA		**Final Model**				
Skill Score/Case	SP	Site	Intcpt	STEP 2	#SpEnc	Female	Variance explained by predictors
Checklist Scores							
Biomedical	2.9%	11.6%	53.0*	.16*	-.03		6.7%
Routine Counseling 1	2.7%	8.3%	71.7*	.09*	.21		23.0%
Grave Illness	2.0%	3.0%	Predictors not added due to small variance proportions				
IPS Scores							
Biomedical	26.5%	0.0%	79.7*		-.05		0.0%
Routine Counseling 1	27.1%	15.7%	78.1*		.01		56.4%
Routine Counseling 2	49.4%	n/a	72.7*		.13	4.08	7.6%
Grave Illness	44.8%	0.0%	76.0*		-.04		0.0%

*Effect is statistically significant at the .05 level.

References

Bryk AS, Raudenbush SW, Congdon R. Hierarchical Linear Modeling with the HLM/2L and HLM/3L Programs. Chicago: Scientific Software International. 1994.

Boulet, J., Friedman, Ben-David, M., Hambleton, R.K., Burdick, W., Ziv, A. & Gary, N.E. (1998). An investigation of the sources of measurement error in the post-encounter note written scores from standardized patient examinations. Advances in Health Sciences Education, *3*(2), 89-100.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Sage Publishing, Newbury Park, Ca.

Colliver, J.A., Marcy, M.L., Vu, N.V., Steward, D.E., Robbs, R.S. (1994). Effect of multiple standardized patients to rate interpersonal and communication skills on intercase reliability. *Teaching and Learning in Medicine 6*(1). 45-58.

Colliver, J.A., Robbs, R.S., Vu, N.V. (1991). Effects of using two or more standardized patients to simulate the same case on case means and case failure rates. *Academic Medicine, 66*(10). 616-618.

DeChamplain, A.F., Gessaroli, M.E., Floreck, L.M. (2000). Assessing the impact of standardized patient variability on examination master level decision consistency rates. A paper presented at the meeting of the National Council on Measurement in Education. New Orleans, LA.

DeChamplain, A.F., Macmillan, M., Klass, D., Margolis, M. (1999). Assessing the impact of intra-site and inter-site checklist recording discrepencies on the reliability of scores

obtained in a nationally administered standardized patient examination. *Academic Medicine, 74,* s52-s54.

DeChamplain, A.F., MacMillan, M.K., Margolis, M.J., King, A.M., Klass, D.J. (1998). Do discrepencies in standardized patients' checklist recording affect case and examination mastery level decision? *Academic Medicine 73*(10). s75-77.

Klass, D., Fletcher, E., King, A., Durinzi, D., Nungester, R.J., Clauser, B. Ripkey, D. (1992). Paper presented at the International Conference Proceding: Approaches to the assessment of clinical competence Part I. Dundee, Scotland.

Petrusa, E.R., Blackwell, T.A., Carline, J., Ramsey, P.G., McGaghie, W., Colindres, R., Kowlowitz, V., Mast, T.A., Soler, N. (1991). A multi-institutional trial of objective structured clinical examination. *Teaching and Learning in Medicine 3*(2) 86-94.

Reznick, R. The Medical Council of Canada's experience with the use of standardized patient examinations for high stakes testing. In: Proceedings from the Eighth International Ottawa Conference. Philadelphia, PA: National Board of Medical Examiners, 2000.

Reznick, R.K., Blackmore, D., Cohen, R., Baumer, J. Rothman, A., Smee, S., Chalmers, A., Poldre, P., Birtwhistte, R., Walsh, P., Spady, D., Berard, M. (1993). An objective structured clinical examination for the licentiate of the Medical Council of Canada: From research to reality. *Academic Medicine 68(10).* s4-6.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer.* Sage Publishing, Newbury Park, Ca.

Swanson, D.B., Norcini, J.J. (1989). Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine. 1*(3). 158-166.

Tamblyn, R.M., Klass, D.J., Schnabl, G.K., & Kopelow, M.L. (1991). Sources of unreliability and bias in standardized-patient rating. *Teaching and Learning in Medicine (3)*2. 74-85.

van der Vleuten, C.P., Swanson, D.B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine 2*(2). 58-76.

Whalen, GP. Educational Commission for Foreign Medical Graduates clinical skills assessment. In: Proceedings from the Eighth International Ottawa Conference. Philadelphia, PA: National Board of Medical Examiners, 2000.